DOCUMENT RESUME

ED 137 357 TM 006 158

TITLE	Computers and Testing: Steps Toward the Inevitable Conquest. Professional Series 76-1.
INSTITUTION	Civil Service Commission, Washington, D.C. Personnel Measurement Research and Development Center.
PUB DATE	Sep 76
NOTE	45p.; A Symposium held at the Annual Meeting of the American Psychological Association (83rd, Chicago, Illinois, August 31, 1975) ; Not available in hard copy due to marginal legibility of original
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Bayesian Statistics; *Computer Oriented Programs; Computer Programs; Guessing (Tests); *Item Analysis; *Item Banks; Mathematical Models; Multiple Choice Tests; Probability; Response Style (Tests); Scoring; Simulation; *Testing; Test Validity
IDENTIFIERS	Bayesian Tailored Testing; Latent Trait Theory; *Tailored Testing; Urry Item Parameterization Technique

ABSTRACT

This pamphlet reprints three papers and an invited discussion of them, read at a Division 5 Symposium at the 1975 American Psychological Association Convention. The first paper describes a Bayesian tailored testing process and shows how it demonstrates the importance of using test items with high discrimination, low guessing probability, and a uniform (rectangular) distribution of difficulty. The second paper describes a relatively easy, inexpensive, and reliable method of estimating these three item parameters. The procedure assumes that the underlying ability is normally distributed. Data were examined to determine the reasonableness of the assumption. The third paper reports the results of two simulation studies evaluating the effectiveness of the parameterization technique, showing that the procedure works well under "standard" conditions (at least 2000 cases and 100 items) and is adequate with 1000 cases and 60 items. The discussion paper points out that accurate estimations of item parameters are essential to successful tailored testing; compares the Bayesian technique with parameterization based on maximum likelihood estimates; and concludes that there is now no apparent obstacle to successful operational tailored testing. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

Personnel Research and Development Center

Professional Series 76-1

Computers and Testing: Steps Toward the Inevitable Conquest

ERRATA

Page 14, column 2, line 3: *For* score *read* scores

Page 15, column 1, equation 3: *Read*

$$\hat{P}_i = (1 - \hat{\hat{c}}_i)^{-1} (\hat{P}_i - \hat{\hat{c}}_i)$$

Page 15, column 2, equation 6: *Read*

$$\hat{\beta}_{I\theta} = \hat{\beta}_{i\theta} (\hat{P}_i \hat{Q}_i)^{\frac{1}{2}} \{(1 - \hat{\hat{c}}_i) \phi (\hat{\hat{Y}}_i)\}^{-1}$$

Page 15, column 2, equation 9: *Read*

$$\chi_i^2 = \sum_{j=0}^{m-1} (r_j - n_j P_i^*(j))^2 (n_j P_i^*(j) Q_i^*(j))^{-1}$$

Page 16, column 1, equation 12: *Read*

$$P_i(j) = \text{Prob}\,(t \geq -a_i(z_j - b_i))$$

$$= (2\pi)^{-\frac{1}{2}} \int_{-a_i(z_j - b_i)}^{\infty} \exp\left[\frac{-t^2}{2}\right] dt$$

Page 17, column 1, equation 15: *Read*

$$c(\hat{\hat{c}}_i) = \left\{ I(\hat{\theta})^{-1} \left[\frac{\partial I(\hat{\theta})}{\partial \hat{\hat{b}}_i} - \frac{\partial^2 I(\hat{\theta})}{\partial \hat{\hat{a}}_i \partial \hat{\hat{c}}_i} \right] \right\}_{\hat{\theta} = \hat{\hat{b}}_i}$$

Page 17, column 1, line 7 from bottom: *For* there is no no need *Read* there is no need

Page 17, column 2, line 12: *For* [- ∞, -2.45], [-2.45, -2.40]....[2.45, ∞] *read* [- ∞, -2.45], (-2.45, -2.40]...(2.45, ∞]

Page 17, column 2, equation 22: *Read*

$$\chi_i^2 = \sum_{j=1}^{k} [r_j - n_j P_i^*(\nu_j)]^2 [n_j P_i^*(\nu_j) Q_i^*(\nu_j)]^{-1}$$

Page 18 should contain figure 1, now on page 33

Page 19, column 1, line 5: *For* The first study examines, *read* The first study (Urry, 1976) examines

Page 19, column 1, line 7: *For* the second explores, *read* the second (Gugel, Schmidt, & Urry, 1976) explores

Page 29, column 2; Add to reference list:

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary estimation procedure. In *Proceedings of the first conference on computerized adaptive testing*. (PS-75-6, Personnel Research and Development Center, U.S. Civil Service Commission) Washington, D.C. U.S. Government Printing Office, 1976. Stock No. 006-000-00940-9

Urry, V. W. A five year quest: is computer-assisted testing feasible? In *Proceedings of the first conference on computerized adaptive testing*. (PS-75-6, Personnel Research and Development Center, U.S. Civil Service Commission) Washington, D.C. U.S. Government Printing Office, 1976. Stock No. 006-000-00940-9

Page 33, Figure 1 should be on page 18

Page 36, column 2, penultimate line: *For* Baysian *read* Bayesian

PS-76-1

COMPUTERS AND TESTING: STEPS TOWARD THE INEVITABLE CONQUEST

A Symposium Held at the 83rd Annual Convention
of the American Psychological Association,
Chicago, Illinois, August 31, 1975

Research Section
Personnel Research and Development Center
U.S. Civil Service Commission
Washington, D.C.
September 1976

COMPUTERS AND TESTING: STEPS TOWARD THE INEVITABLE CONQUEST

ABSTRACT

This pamphlet reprints three papers and an invited discussion of them, read at a Division 5 Symposium at the 1975 APA Convention. The first paper describes a Bayesian tailored testing process and shows how it demonstrates the importance of using test items with high discrimination, low guessing probability, and a uniform (rectangular) distribution of difficulty. The second paper describes a relatively easy, inexpensive, and reliable method of estimating these three item parameters. The procedure assumes that the underlying ability is normally distributed. Data were examined to determine the reasonableness of the assumption. The third paper reports the results of two simulation studies evaluating the effectiveness of the parameterization technique, showing that the procedure works well under "standard" conditions (at least 2000 cases and 100 items) and is adequate with 1000 cases and 60 items. The discussion paper points out that accurate estimations of item parameters are essential to successful tailored testing; compares the Bayesian technique with parameterization based on maximum likelihood estimates; and concludes that there is now no apparent obstacle to successful operational tailored testing.

CONTENTS

	Page
Opening Remarks	
William A. Gorham, U.S. Civil Service Commission	1
Bayesian Tailored Testing and the Influence of Item Bank Characteristics	
Carl J. Jensema, Gallaudet College	3
Ancillary Estimators for the Item Parameters of Mental Test Models	
Vern W. Urry, U.S. Civil Service Commission	14
The Urry Item Parameter Estimation Technique: How Effective?	
Frank L. Schmidt and John F. Gugel, U.S. Civil Service Service Commission	19
Discussion:	
Frederic M. Lord, Educational Testing Service	31

OPENING REMARKS

William A. Gorham
Chairperson

Over five years ago, the man who is perhaps the foremost theoretician and researcher in the area of tailored testing listed five major technical limitations on tailored testing applications (Lord, 1970):

(1) Test items may not group into subtests (or item banks) measuring substantially different psychological dimensions.

(2) Items differ, in practice, on discriminatory power (a_i) and the guessing parameter (c_i), as well as on difficulty level (b_i). This greatly complicates the problem of determining the optimal sequence of item administration; based on intermediate (provisional) estimates of standing on the latent trait, θ. No such optimal procedure exists.

(3) Accurate estimation of the item parameters a_i, b_i, and c_i is absolutely necessary for tailored testing, but is at present a "difficult and expensive, and hazardous operation."

(4) Accuracy is particularly important in the case of item difficulty, b_i. "If there is any doubt about the accuracy of the estimated values of item difficulties b_i, there will be doubt about the accuracy and fairness of final scores given to examinees" (Lord, 1970, p. 180).

(5) Better measurement than that provided by tailored testing can usually be attained by selecting, for example, the best 60 items (highest a_i) and presenting these as a conventional test. *"This may actually prove to be a fatal objection to any general use of tailored testing"* (Lord, 1970, p. 180; emphasis as in original).

The papers in this symposium directly address four of these five problems and provide answers and solutions that are the best advanced to date. Addressing the second problem listed above, the first paper describes a computer-based Bayesian procedure for optimal sequential selection of items to be administered. This procedure takes into account simultaneously the a_i, b_i, and c_i values of all items in the item bank and the examinee's current ability estimate on the latent trait, $\hat{\theta}$. Results of simulation studies are presented which indicate the procedure is highly efficient. The second paper provides a solution for the third and fourth problems listed by Lord (1970). A method of estimating the item parameters a_i, b_i, and c_i is presented which is less than one percent as costly as the commonly used maximum likelihood procedure. This paper also presents the results of simulation studies which indicate the method is highly accurate in the case of all three parameters. The derivations basic to this item parameterization procedure indicate a solution to Lord's (1970) fifth -- and potentially most serious -- problem. Estimates of attainable item discriminatory values (a_i) available in 1970 were based on sample statistics (usually item inter-r's) attenuated by guessing. Thus they led to underestimations of attainable a_i values. This conclusion from the derivation has since been substantiated by $\hat{a}_i$'s obtained on large sample empirical data at the U.S. Civil Service Commission. The final paper presents results of a simulation study showing that item parameter estimates derived from the item parameterization procedure described in the preceding paper, when used

in tailored testing, perform essentially as well as the actual parameters.

Thus the symposium directly addresses and provides solutions for four of the five major technical difficulties standing in the way of actual practical application of the tailored testing concept.

In view of this, the U.S. Civil Service Commission is moving toward the implementation by 1979 - 80 of a nationwide computer-assisted testing system for Federal jobs. The impact of this major advance on other large-scale testing programs is certain to be significant.

BAYESIAN TAILORED TESTING AND THE INFLUENCE OF ITEM BANK CHARACTERISTICS*

Carl J. Jensema

Gallaudet College

Conventional tests are generally constructed to discriminate over a rather wide range of examinee ability. One of the consequences of this approach is that a conventional test usually contains many items which are not appropriate for a particular level of ability. Psychometricians have long been aware of this and in recent years they have increasingly turned their attention to the possibility of programming computers to design and administer tests.

Of the many computerized testing methods which have been proposed, the Bayesian process developed by Owen (1969) seems to be the most elegant and intuitively appealing method. It assumes locally independent binarily scored items and a normal ogive model (Lord and Novick, 1968, Ch. 16) in which the probability of passing a free response item g at ability level θ is expressed as

$$(1) \qquad P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} \exp\left[\frac{-t^2}{2}\right] dt$$

If the item is not of the free response type and c_g is the probability of guessing correctly, the probability of passing becomes

$$(2) \qquad P'_g(\theta) = P_g(\theta) + c_g[1 - P_g(\theta)]$$

The derivation of Owen's Bayesian tailoring process has been described several times in the literature (Owen, 1969; Urry, 1971; Jensema, 1974a). We will briefly run through the fundamental formulas here for the sake of completeness.

Suppose $N(\theta_0, \sigma_0^2)$ expresses our knowlege of an examinee having ability θ. If we administer free response item g, which has discrimination and difficulty parameters a and b, and if the examinee responds correctly, Bayes's theorem specifies that the information available is

$$(3) \qquad P(\theta|1) = k\, P_g(\theta)\, (\sqrt{2\pi}\,\sigma_0)^{-1} \exp\left[\frac{-(\theta - \theta_0)^2}{2\sigma_0^2}\right]$$

where $P_g(\theta)$ is defined by (1) and k is such that

$$(4) \qquad \int_{-\infty}^{\infty} P(\theta|1)\, d\theta = 1 \ .$$

The solution is

$$(5) \qquad k^{-1} = 1/2\,(1 - \operatorname{erf} D)$$

where $erf\ D$ is the error function

$$(6) \qquad \operatorname{erf} D = \frac{2}{\sqrt{\pi}} \int_0^D \exp(-t^2)\, dt$$

and

$$(7) \qquad D = \frac{b - \theta_0}{\sqrt{2(a^{-2} + \sigma_0^2)}}$$

The expectation of the posterior mean is

---

*This research was supported by the Office of Demographic Studies, Gallaudet College, Washington, D.C., 20002.

(8) $$E(\theta|1) = \theta_o + \frac{\sqrt{2}\,\sigma_o^2}{\sqrt{\pi\,(a^{-2} + \sigma_o^2)}} \exp(-D^2)\,(1 - \operatorname{erf} D)^{-1}$$

and the variance is

(9) $$\operatorname{var}(\theta|1) = \sigma_o^2 \left[1 - \frac{\frac{2}{\sqrt{\pi}} - 2D \exp(D^2)\,(1 - \operatorname{erf} D)}{\sqrt{\pi}(1 + a^{-2}\sigma_o^{-2})\,(\exp(D^2)\,(1 - \operatorname{erf} D))^2} \right].$$

Similarly, if the examinee gives a wrong response to item g we have

(10) $$P(\theta|0) = \frac{k}{k-1}\,(1 - P_g(\theta))\,(\sqrt{2\pi}\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2\sigma_o^2} \right],$$

(11) $$E(\theta|0) = \theta_o - \frac{\sqrt{2}\,\sigma_o^2}{\sqrt{\pi(a^{-2} + \sigma_o^2)}} \exp(-D^2)\,(1 + \operatorname{erf} D)^{-1},$$

and

(12) $$\operatorname{var}(\theta|0) = \sigma_o^2 \left[1 - \frac{\frac{2}{\sqrt{\pi}} - 2D \exp(D^2)\,(1 + \operatorname{erf} D)}{\sqrt{\pi}\,(1 + a^{-2}\sigma_o^{-2})\,(\exp(D^2)\,(1 + \operatorname{erf} D))^2} \right].$$

To expand this discussion a little further, assume that item g is not a free response item and that it has a probability of C_g of guessing correctly. If the examinee gives a correct response we have

(13) $$P'(\theta|1) = \lambda\,P_g'(\theta)\,(\sqrt{2\pi}\,\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2\sigma_o^2} \right],$$

(14) $$E'(\theta|1) = \theta_o + (1 - C_g)\,k^{-1}\lambda S,$$

and

(15) $$\operatorname{var}'(\theta|1) = \sigma_o^2 - (1 - C_g)\,k^{-1}\lambda S^2\,(t - C_g\lambda)$$

where the prime is used to signify the effect of guessing, $P_g'(\theta)$ is defined by (2), and we take

(16) $$\lambda^{-1} = C_g + (1 - C_g)\,k^{-1},$$

(17) $$S = k\,\sigma_o \exp(-D^2)\,(2\pi(1 + a^{-2}\sigma_o^{-2}))^{-\frac{1}{2}}$$

and

(18) $$t = 1 - 2\sqrt{\pi}\,k^{-1} D \exp(D^2).$$

If the examinee gives a wrong response, the formulas in (10), (11), and (12) hold, since our information, that the examinee does not know the correct answer, is the same as in the free response case.

Now assume we have n items and want to select the best one for administration. The expected posterior variance of θ after administration of a particular item is

$$E_u(\operatorname{var}(\theta|u)) = \theta_o^2 + \sigma_o^2 - P(0)\,[E(\theta|0)]^2 - P(1)\,[E(\theta|1)]^2$$

(19) $$= \sigma_o^2 \left[1 - \frac{2}{\pi\,(1 + a^{-2}\sigma_o^{-2})\exp(2D^2)\,(1 - (\operatorname{erf} D)^2)} \right]$$

when items are of the free response type and

(20) $$E'_u(\operatorname{var}(\theta|u)) = \sigma_o^2 \left[1 - \frac{(1 - C_g)\,\lambda\,(1 + C_g\,(1 - k^{-1}))}{2\pi\,(1 + \sigma_o^{-2}a^{-2})\,(1 - k^{-1})\exp(2D^2)} \right]$$

when the items are affected by guessing. In (19) and (20) u refers to the correctness of the examinee's response and is taken as 1 or 0. The item which leads to the smallest expected posterior variance is the most desirable one to administer. It is sufficient to select the item with the smallest value α where

(21) $$\alpha = (a^{-2} + \sigma_o^2)\,(1 - (\operatorname{erf} D)^2)\exp(2D^2)$$

for free response items and

(22) $$\alpha' = \left(\frac{1}{1 - C_g}\right)(1 + \sigma_o^{-2}a^{-2})\,(1 - k^{-1})\,\lambda^{-1}\exp(2D^2),$$

when guessing is present.

If we have a pool of n items and estimates of the normal ogive model parameters

for each item, we may use a Bayesian sequential procedure to select items for administration to a particular examinee. Let $\hat{\theta}(m)$ and $\hat{\sigma}^2(m)$ be an estimate of the examinee's ability and its variance where m indicates the number of items administered. Assume the population has ability distributed as $N(0,1)$ and take $\hat{\theta}(o)$ and $\hat{\sigma}(o)$ as 0 and 1. Calculate α_i values for all (unused) items, $i = 1, 2, \ldots, (n - m)$, using (22). (We will assume that the items are not free-response). The examinee is administered the item with the smallest α_i value. If an incorrect response is given, $\hat{\theta}(m + 1)$ and $\hat{\sigma}^2(m + 1)$ are calculated from (11) and (12). If the response is correct, (14) and (15) are used. This cycle is repeated until $\hat{\sigma}(m)$ is within some pre-selected limit. The selection of a $\hat{\sigma}(m)$ value for termination is, of course, arbitrary. It is usually selected to yield some expected level of validity according to

$$r_{\theta\hat{\theta}} = \sqrt{1 - \hat{\sigma}^2(m)} \quad . \tag{23}$$

The characteristics of an item bank used for tailored testing are very important to the efficiency and accuracy of the process. There are four basic requirements for a good item bank. These have been mentioned in whole or part in a number of publications (i.e., Urry, 1970, 1971a, 1971b, 1974; Jensema, 1972, 1974a, 1974b; etc.) and may be summarized as follows:

1. Item discrimination should be as high as possible and should not be less than .8.
2. Item guessing probabilities should be as low as possible.
3. The item bank must consist of a sufficiently large number of items.
4. Item difficulties should have a rectangular distribution.

The remainder of this paper will concentrate on demonstrating the importance of each of these four requirements.

Assume that an infinitely large item bank exists and that all items have the same discriminatory power and the same probability of guessing correctly. The assumption of an infinitely large item bank allows the selection of an item i having a difficulty level exactly equal to any given estimate of ability. When this can be done many of the formulas may be greatly simplified since we have:

$$D_i = 0 \tag{24}$$

and

$$\operatorname{erf} D_i = 0 \, . \tag{25}$$

The equations for $\hat{\sigma}_{(m+1)}^2$ for correct and incorrect responses become

$$\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2a_i(1-c_i)^2}{\pi(1-c_i)^2} \right] \tag{26}$$

and

$$\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2a_i}{\pi} \right] \tag{27}$$

where m is the number of items previously administered.

An items i's difficulty is the point at which the probability of knowing the correct answer is exactly .5. If guessing is in effect, the probability of responding correctly is equal to the probability of knowing the answer plus the probability of guessing correctly. Then $\hat{\sigma}_{(m+1)}^2$ may be expected to be the sum of (26) and (27) weighted by the probabilities of a correct or incorrect response:

$$\mathcal{E}\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[.5(1+c_i)\left(1 - \frac{2a_i(1-c_i)^2}{\pi(1+c_i)^2}\right) + .5(1-c_i)\left(1 - \frac{2a_i}{\pi}\right)\right] \tag{28}$$

A little algebraic manipulation reduces this to

$$\mathcal{E}\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2a_i(1-c_i)}{\pi(1+c_i)} \right] . \tag{29}$$

Inserting appropriate values for a_i and c_i in equation (29) and plotting the results against the number of items administered demonstrates the influence of item discrimination and guessing probability on the tailoring process. Figure 1 plots the expected standard error of the estimate $\mathcal{E}\hat{\sigma}(m+1)$ by the number of items administered for five levels of discrimination when

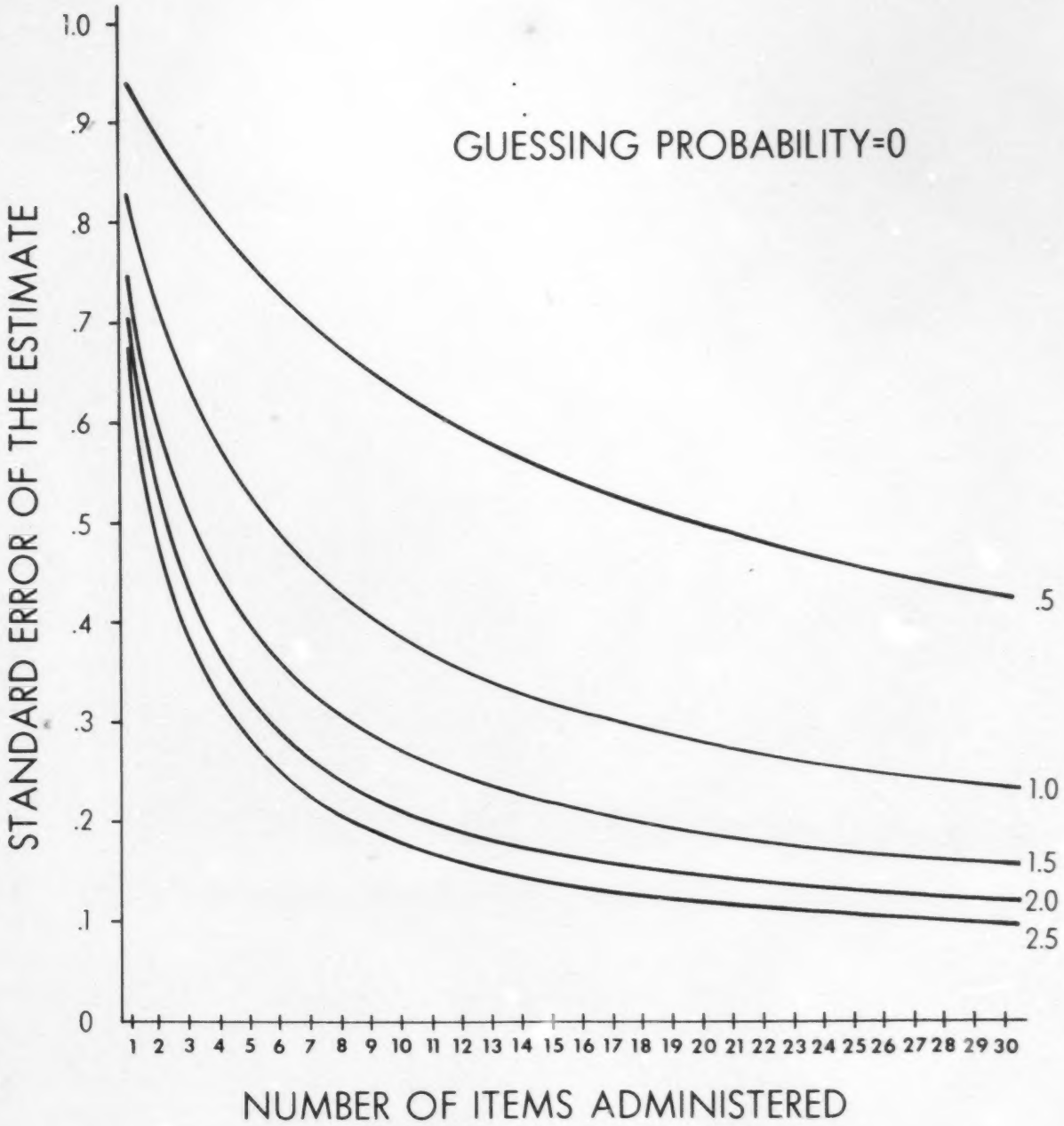


Figure 1. Expected standard error of the estimate according to number of items administered at five levels of item discrimination.

guessing probability is zero and an infinite number of items are available. Notice the sharp difference in the number of items needed at different levels of discrimination. For example, if the items have discriminatory powers of 2.5 only 4 or 5 items are needed to reach a standard error of the estimate of .30, while 17 or 18 items are needed to reach this level when item discrimination is only 1.0.

Now suppose we take item discrimination to be 1.0, a rather low value which is easily obtained. Figure 2 plots the expected standard error of the estimate for various guessing values by the number of items administered. The guessing values range from .5 (i.e., true-false items) to 0.0 (i.e., free response items). The greater the probability of guessing, the more items required to reach a specific standard error of the estimate.

To give a clear example of the combined effects of discrimination and guessing on the tailoring process, suppose we have three item banks which, for convenience, are referred to as I, II, and III. Assume Bank I items have discrimination and guessing parameters of .5 and .33. Bank II's parameters are 1.0 and .25 while Bank III has parameter values of 2.0 and .20. These banks may be roughly classified as unacceptable, fair, and excellent for tailored testing purposes. Assuming that each bank has an infinite number of items, and plotting the expected standard error of the estimate against the number of items administered, the three curves in Figure 3 are obtained.

In Figure 3, notice that Bank I would give unacceptable results. After 30 items the expected standard error of the estimate is only .56 (i.e., reliability = .69, validity = .83). In contrast an excellent item bank, such as Bank III, would reach this level after only 3 or 4 items. The advantage of high discrimination and low guessing probability in an item bank is obvious.

Up to this point we have discussed the behavior of Bayesian tailored testing when the item bank is assumed to be of unlimited size. The obvious question which follows is what happens when item bank sizes are within practical limits? To answer this question, Monte Carlo data for 200 items are generated for each of 100 "examinees" using Urry's (1970) "LOGIST" program. The parameters for discrimination (1.0) and guessing (.25) were the same as for Bank II mentioned earlier. Eight sets of 25 difficulty values (-2.4, -2.2, ..., 0.0, ..., 2.2, 2.4) were employed. Bayesian tailored testing was simulated with this data using 50, 75, 100, 150, and 200 items in the bank. Since difficulty had been specified in sets of 25 values, the item banks had 2, 3, 4, 6, and 8 items at each of the 25 difficulty levels respectively.

For each of the five item banks and for each of the 100 examinees, tailoring was simulated until 30 items had been "administered." As each item was "administered" the new estimate of ability was recorded. Since the data were randomly generated, true ability [distributed as $N(0,1)$] was known and could be correlated with estimated ability. Table 1 gives the validity (correlation between true and estimated ability) for each item bank by the number of items "administered." The last column in Table 1 gives the expected validities for an item bank of infinite size as calculated from equations (22) and (23).

The Monte Carlo data above represents items which are passable but not especially good for tailored testing. To see how item bank size would influence validity when the bank was composed of excellent items, the Monte Carlo data tailoring simulation was repeated with higher discrimination (2.0) and lower guessing (.20) parameter values. These configurations correspond to Bank III mentioned earlier. The results of the simulated tailoring with this new data are given in Table 2.

For practical application it is apparent that a very large number of items is not a critical item bank characteristic if the bank is good in other respects. In both Table 1 and Table 2, the Monte Carlo data validities obtained for the five banks closely match each other and they also parallel the validities to be expected from a corresponding item bank of infinite size. However, it must be remembered that this was Monte Carlo data and the tailoring simulation used known parameter values for discrimination, difficulty, and guessing. With real data involving imprecise parameter estimates and a possible non-uniform distribution of difficulty, it would be wise to be a bit cautious if a bank had, say, less than 75 items. In connection with this, there are some practical problems which arise if an item bank is too large. A large bank has more items available for administration, but

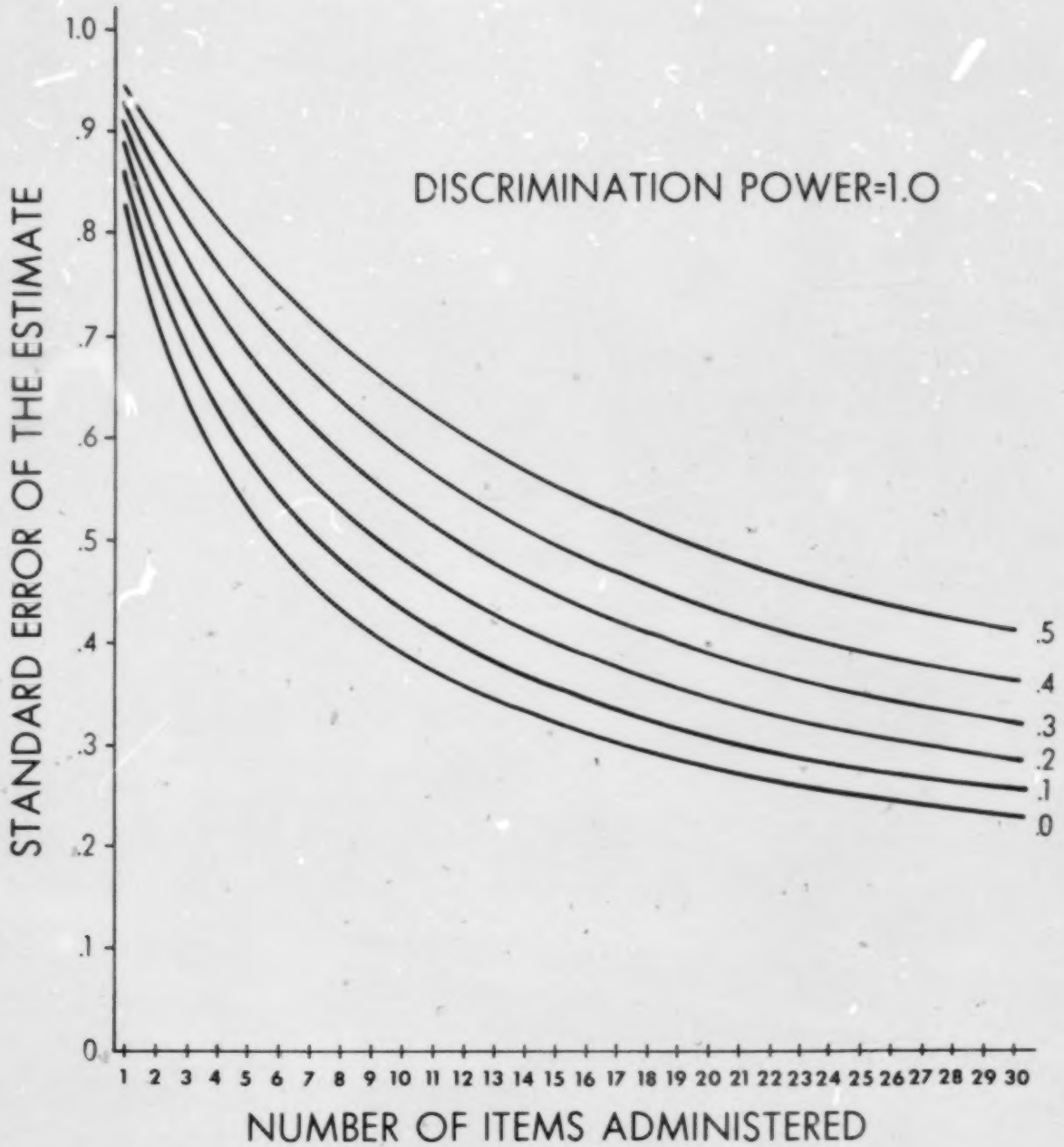


Figure 2. Expected standard error of the estimate according to number of items administered at six guessing probabilities.

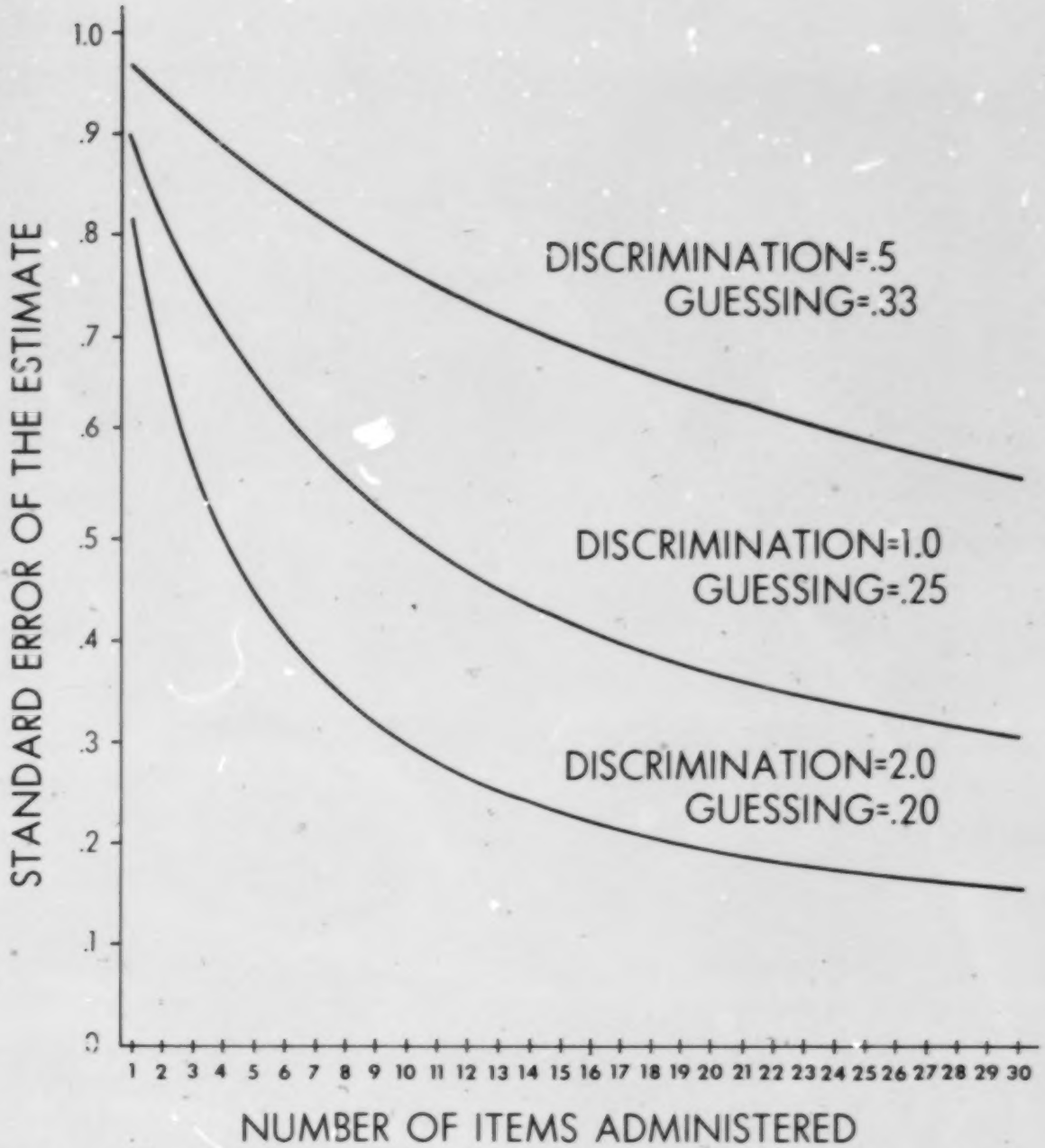


Figure 3. Expected standard error of the estimate for three item banks according to number of items administered.

TABLE I

VALIDITY ($r_{\theta\hat{\theta}}$) OBTAINED WITH DIFFERENT SIZE ITEM BANKS

(MONTE-CARLO DATA, N=100, A=1.0, C=.25)

ITEMS ADMINISTERED	ITEMS IN BANK					
	50	75	100	150	200	∞*
1	.53	.53	.53	.53	.53	.44
2	.59	.59	.59	.59	.59	.57
3	.65	.65	.65	.65	.65	.66
4	.72	.72	.72	.72	.72	.72
5	.78	.78	.78	.78	.78	.76
6	.81	.80	.80	.80	.80	.79
7	.83	.82	.82	.82	.82	.81
8	.84	.84	.84	.84	.84	.83
9	.85	.85	.84	.84	.84	.85
10	.86	.86	.86	.85	.85	.86
11	.86	.87	.88	.87	.87	.87
12	.87	.87	.89	.87	.87	.88
13	.89	.89	.89	.87	.88	.89
14	.90	.91	.90	.88	.88	.90
15	.91	.91	.91	.90	.90	.91
16	.91	.92	.92	.91	.91	.91
17	.92	.92	.92	.92	.91	.92
18	.92	.92	.93	.92	.92	.92
19	.92	.92	.93	.92	.92	.93
20	.93	.93	.93	.93	.93	.93
21	.93	.93	.93	.93	.93	.93
22	.93	.94	.94	.94	.93	.94
23	.93	.94	.94	.94	.94	.94
24	.93	.94	.94	.94	.94	.94
25	.93	.94	.95	.94	.94	.94
26	.94	.95	.95	.94	.94	.95
27	.94	.95	.95	.94	.95	.95
28	.94	.95	.95	.95	.95	.95
29	.94	.95	.95	.95	.95	.95
30	.94	.95	.95	.95	.95	.95

* Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items

TABLE 2

VALIDITY ($r_{\theta\hat{\theta}}$) OBTAINED WITH DIFFERENT ITEM BANK SIZES

(MONTE-CARLO DATA, N=100, A=2.0, C=.2)

ITEMS ADMINISTERED	ITEMS IN BANK					
	50	75	100	150	200	∞*
1	.66	.66	.66	.66	.66	.58
2	.75	.75	.75	.75	.75	.74
3	.84	.84	.84	.84	.84	.82
4	.89	.89	.89	.89	.89	.86
5	.92	.92	.92	.92	.92	.90
6	.93	.93	.93	.93	.93	.91
7	.94	.94	.94	.94	.94	.93
8	.95	.95	.95	.95	.95	.94
9	.96	.95	.95	.95	.95	.95
10	.96	.96	.96	.96	.96	.96
11	.97	.96	.96	.96	.96	.96
12	.97	.96	.96	.96	.97	.96
13	.97	.97	.97	.97	.97	.97
14	.97	.97	.97	.97	.97	.97
15	.97	.97	.98	.97	.98	.97
16	.97	.98	.98	.98	.98	.98
17	.97	.98	.98	.98	.98	.98
18	.98	.98	.98	.98	.98	.98
19	.98	.98	.98	.98	.98	.98
20	.98	.98	.98	.98	.98	.98
21	.98	.98	.98	.98	.98	.98
22	.98	.98	.99	.98	.98	.98
23	.98	.98	.99	.98	.98	.98
24	.98	.98	.99	.98	.98	.98
25	.98	.98	.99	.99	.99	.98
26	.98	.98	.99	.99	.99	.99
27	.98	.98	.99	.99	.99	.99
28	.98	.98	.99	.99	.99	.99
29	.98	.98	.99	.99	.99	.99
30	.98	.98	.99	.99	.99	.99

* Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items

the storage requirements and the increased computer processing needed for item selection also slow things down while adding to overall computer costs. (Some good cost-efficiency studies are needed on this!)

The last item bank requirement is uniform distribution of difficulty. The exact results of violating this rule are difficult to predict, since they would necessarily depend on the actual distribution of item difficulty, the discrimination and guessing parameter values, the number of items in the bank, and the criteria used to terminate the tailoring process. The essential point to remember is that the Bayesian tailoring procedure attempts to select for administration the item which will yield the most information. If at a particular level of difficulty, there are no items available, the Bayesian process will be forced to select an item which is not appropriate and which will yield less than an optimal amount of information.

One issue which is often overlooked in considering the distribution of item difficulties is the possible inaccuracy of parameter estimates at extreme levels of difficulty. Consider that item parameter-estimates are based on the responses of a population assumed to have ability θ distributed as $N(0,1)$. Further, consider that an item's difficulty is defined as the ability level at which the probability of knowing the correct answer is exactly .5. The number of examinees with ability at or near the true difficulty level of an item is critical for accurate parameter estimation. If the true difficulty of an item is, say, 3.0 or some other extreme value and if θ is normally distributed, it will take a very, very large sample size to provide a sufficient basis for obtaining accurate parameter estimates for the item. If the parameter estimates are inaccurate, the use of them in the Bayesian tailoring sequence is obviously not going to lead to satisfactory results.

To date, there has been no systematic study of the effect of sample size on parameter estimation for items having extreme levels of difficulty. Such a study could become quite complex, especially if the investigation was carried over into the comparative sensitivity of various estimation techniques. For practical purposes, at this stage in the art of item bank development, it may be wise to routinely discard items having difficulty estimates of less than -3.0 or greater than 3.0, and to look with some suspicion on the remaining items having estimated difficulties outside the -2.5 to 2.5 range.

To summarize, this paper has outlined a Bayesian approach to item selection for tailored testing. Four basic requirements of a good item bank for this process have been discussed. If these requirements are met, Bayesian tailored testing will yield excellent results. They key to the process lies in careful construction of item banks. If attention is given to this, the Bayesian tailoring process gives us a fundamental tool for practical application of latent trait mental test theory.

REFERENCES

Jensema, C. J. *An application of latent trait mental test theory to the Washington pre-college testing battery.* Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C. J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 29 - 48. (a)

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement*, 1974, 34, 757 - 766. (b)

Lord, F. M., and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Owen, R. J. A Bayesian approach to tailored testing. *Research Bulletin*, 69-92. Princeton, N.J.: Educational Testing Service, 1969.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished doctoral dissertation. Purdue University, 1970.

Urry, V. W. Approximation methods for the item parameters of mental test models. *Research Bulletin*, 0871-202. Seattle: University of Washington, Bureau of Testing, 1971. (a)

Urry, V. W. Individualized testing by Bayesian estimation. *Research Bulletin*, 0171-177. Seattle: University of Washington, Bureau of Testing, 1971. (b)

Urry, V. W. Approximations of item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, 34, 253 - 269.

ANCILLARY ESTIMATORS FOR THE ITEM PARAMETERS OF MENTAL TEST MODELS

Vern W. Urry

Personnel Research and Development Center
U.S. Civil Service Commission

Five years ago it was observed that the estimation of item parameters required by computer assisted testing was "a difficult, expensive, and hazardous operation (Lord, 1970, p. 180)." My purpose today is to outline a new procedure for the estimation of item parameters that is neither difficult, nor expensive, nor hazardous. The procedure is applicable in the number of instances where the distribution of latent ability is reasonably well approximated by a normal distribution. Under this assumption, the accuracy of the procedure has been studied through simulation techniques. The results of this research have illustrated the merit of the procedure and are reported in the subsequent paper.

A program for the procedure is currently available for use on various computers, and we would be happy to furnish source listings and card decks to interested researchers. The program is relatively easy and inexpensive to execute, requiring only a few minutes of central processing time for large amounts of data.

The output of the program is an item analysis yielding ancillary estimates for a_i, item discriminatory power; b_i, item difficulty; and c_i, item coefficient of guessing. In overview, estimates of the parameters a_i, b_i, and c_i are obtained by an iterative minimum χ-square procedure. The procedure consists of two stages that differ only with respect to the particular measure used for manifest ability. In the first stage, the distribution of manifest ability is represented by corrected raw scores where the item being parameterized is omitted from the scoring. In the second stage, the distribution of manifest ability is represented by Bayesian modal estimates of ability (Samejima, 1969). Generally, Bayesian modal estimates more closely approximate the distribution of latent ability than does the distribution of corrected raw score. Therefore, the second stage constitutes a refinement on the first stage. In both stages the procedure iterates item by item through trial values for c_i to obtain pairs of a_i and b_i consistent with large sample estimates of the item-manifest ability point-biserial correlation and the item p-value. This allows the generation of various item characteristic curves (ICC's). The ICC's are then compared with the regression of the binary item on manifest ability. The ICC that best fits this regression, as indicated by the minimum χ-square, is given by the set of approximations, $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$. The approximations are then corrected for characteristics of the particular sample of items being parameterized to obtain "ancillary estimates", $\tilde{a}_i$, $\tilde{b}_i$, and $\tilde{c}_i$. Ancillary estimation as a method was advocated by Fisher (1950). It involves the use of supplementary information from the sample to increase the efficiency of estimators. In the current application the information function or the reciprocal of the error variance of $\hat{\theta}$ at point $\hat{b}_i$ provides the supplementary information required to increase the efficiency of the estimators. The ancillary corrections are necessary because the ICC's are being fitted not to θ, latent ability, but to $\hat{\theta}$, manifest ability. As a result dependencies are induced between the estimates of the independent parameters a_i, b_i, and c_i. For example, it is readily seen that estimates of a_i would be attenuated by error in $\hat{\theta}$ at point $\hat{b}_i$. Indeed, if corrections are not made, $\hat{a}_i$ will be correlated with $\hat{b}_i$; and the correlation need not be linear. Low values of $\hat{a}_i$ may occur with low values of $I(\hat{\theta})_{\hat{\theta}=\hat{b}_i}$ due to the attenuation implicit in low values of $I(\hat{\theta})_{\hat{\theta}=\hat{b}_i}$.

Given adequate numbers of items and cases, the corrections have been found in simulation studies to: (1) decrease the dependencies between the estimates, i.e., $\hat{a}_i$ and $\hat{b}_i$, $\hat{a}_i$ and $\hat{c}_i$, and $\hat{b}_i$ and $\hat{c}_i$; (2) decrease the root-mean-square errors of the form $\left[\sum_{i=1}^{n} (\hat{a}_i - a_i)^2 \, n^{-1}\right]^{\frac{1}{2}}$; $\left[\sum_{i=1}^{n} (\hat{b}_i - b_i)^2 \, n^{-1}\right]^{\frac{1}{2}}$; and $\left[\sum_{i=1}^{n} (\hat{c}_i - c_i)^2 \, n^{-1}\right]^{\frac{1}{2}}$; and (3) increase the correlations between estimated and known parameters, i.e., $r_{\hat{a}a}$, $r_{\hat{b}b}$, and $r_{\hat{c}c}$.

Procedure

From previous developments (Urry, 1974a), we have as the proportion passing a multiple choice item,

$$(1) \qquad P_i' = c_i + (1 - c_i)P_i$$

where the prime indicates that guessing affects the usual item parameter and

$$(2) \qquad P_i = (2\pi)^{-\frac{1}{2}} \int_{\gamma_i}^{\infty} exp\left[\frac{-t^2}{2}\right] dt$$

provides the proportion passing the item if unaffected by guessing. The item coefficient of guessing, c_i, is the lower asymptote of the ICC; and γ_i is the point of cut on the normally distributed variable t assumed basic to the binary response. When unaffected by guessing, a binary response of zero is obtained by the proportion below γ_i; whereas, a binary response of one is obtained by the proportion above γ_i on this hypothetical item continuum.

Given a sample estimate of P_i', the proportion passing an item, we may solve for $\hat{P}_i$ by direct substitution through

$$(3) \qquad \hat{P}_i = (1 - \hat{c}_i)^{-1}(\hat{P}_i' - \hat{c}_i)$$

$$.00 \leq \hat{P}_i \leq 1.00$$

if $\hat{c}_i$ is provisionally assumed. We may solve for $\hat{\gamma}_i$ through equation (2). The point of cut, $\hat{\gamma}_i$, is merely the baseline value in the standard normal distribution above which the proportion $\hat{P}_i$ occurs. The ordinate at the point of cut, $\phi(\hat{\gamma}_i)$ is also needed for subsequent computation.

Again from the same source (Urry, 1974a), we have as the point-biserial correlation between an item subject to guessing and latent ability, θ,

$$(4) \qquad \rho'_{i\theta} = (1 - c_i)\rho_{I\theta}\,\phi(\gamma_i)(P_i'Q_i')^{-\frac{1}{2}}$$

where $\rho_{I\theta}$ is the biserial correlation, unaffected by guessing, between that item and latent ability and the proportion missing the item is given by

$$(5) \qquad Q_i' = 1 - P_i'$$

After the substitution of sample estimates for $\rho'_{i\theta}$, $\phi(\gamma_i)$, P_i', and Q_i' only $\rho_{I\theta}$ remains unknown in equation (4) when $\hat{c}_i$ is provisionally assumed. The solution for $\hat{\rho}_{I\theta}$ is, then,

$$(6) \qquad \hat{\rho}_{I\theta} = \hat{\rho}_{i'\theta}\,(\hat{P}_i'\hat{Q}_i')^{\frac{1}{2}}\{(1 - \hat{c}_i)\phi(\hat{\gamma}_i)\}^{-1}$$

$$.00 \leq \hat{\rho}_{I\theta} \leq 1.00$$

Now from the normal ogive model (Lord, 1968), we have the following definitions for item discriminatory power

$$(7) \qquad a_i \equiv \rho_{I\theta}\,(1 - \rho_{I\theta}^2)^{-\frac{1}{2}}$$

and for item difficulty

$$(8) \qquad b_i \equiv \gamma_i\,\rho_{I\theta}^{-1}.$$

Sample approximations of a_i and b_i are obtained from equations (7) and (8) by substitution of $\hat{\rho}_{I\theta}$ and $\hat{\gamma}_i$ for the corresponding parameters.

As a means of approximating c_i and simultaneously a_i and b_i, we assume provisional values for $\hat{c}_i$ to evaluate

$$(9) \qquad \chi_i^2 = m\text{-}1 \sum_{j=0} \{r_j - n_jP_i'(j)\}^2 \{n_jP_i'(j)Q_i'(j)\}^{-1}$$

Specifically, we seek a minimum for equation (9). The r_j are the number of cases correctly responding to item i at scores of j. The n_j are the number of cases obtaining a given score of j where $j = 0, 1, \ldots, m\text{-}1$ and m is both the number of items and the number of possible discrete

adjusted total test scores when item i is omitted from the scoring. Indeterminate terms, i.e., where $n_j = 0$, are not summed because they merely represent missing data. The $P_i'(j)$ and $Q_i'(j)$ are the expected proportions of correct and incorrect responses to item i among those earning scores of j. These proportions are given by

$$(10)\quad P_i'(j) = c_i + (1 - c_i)\, P_i\,(j)$$

and

$$(11)\quad Q_i'(j) = 1 - P_i'\,(j)$$

In equation (10), the expected proportion correct when unaffected by guessing is given by

$$(12)\quad P_i(j) = Prob\,\{t \geq -a_i(z_j - b_i)\}$$

$$= (2\pi)^{-\frac{1}{2}} \int_{-a_i(z_i - b_i)}^{\infty} exp\left[\frac{-t^2}{2}\right] dt$$

where z_j is the standardized score corresponding to the score of j. Solutions for equation (10) and subsequently equations (11) and (9) are obtained by evaluating the expression $-\hat{a}_i(\hat{z}_j - \hat{b}_i)$ and determining the area or proportion above $-\hat{a}_i(\hat{z}_j - \hat{b}_i)$ as indicated in Equation (12).

Notice the relationship between equation (9) and equations (3), (6), (7), and (8). Given the various provision values of $\hat{c}_i$, $\hat{a}_i$, and $\hat{b}_i$ are fixed by the sample estimates of P_i' and $\rho'_{i\theta}$. Let us illustrate. For example, if $\hat{P}_i'$ equals .761 and $\hat{\rho}'_{i\theta}$ equals .562, we then have

$\hat{c}_i$ (Provisional)	$\hat{a}_i$	$\hat{b}_i$
.03	1.26	-.87
.07	1.33	-.81
.11	1.41	-.75
.15	1.52	-.69
.19	1.67	-.63
.23	1.87	-.56
.27	2.19	-.49

as may be verified through the use of equations (3), (6), (7), and (8). Equation (9) may then be evaluated for each trial set of $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$. The particular trial set of $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ providing the minimum leads to a refinement in the provisional $\hat{c}_i$ in the region of this unrefined minimum. Let us say the minimum value of χ_i^2, the output of equation (9), was located at a provisional $\hat{c}_i$ of .07. As a consequence, we subsequently have

$\hat{c}_i$ (Provisional)	$\hat{a}_i$	$\hat{b}_i$
.04	1.27	-.86
.05	1.29	-.84
.06	1.31	-.83
.07	1.33	-.81
.08	1.35	-.80
.09	1.37	-.78
.10	1.39	-.77

where the $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ that provide a minimum for equation (9) are the approximations for a_i, b_i, and c_i. The search for a minimum equation (9) may be confined to provisional $\hat{c}_i$ on the interval from .00 to .30 since items influenced to a greater extent by guessing are of no interest.

In equation (9), the standardized scores, with the current item omitted, have been substituted for latent ability. Therefore, when a minimum has been found for equation (9), we have minimized the squared standardized discrepancies around the curvilinear regression of a binary item on a substitute for latent ability.

Improved estimators are obtained from the approximations through ancillary corrections. The corrections increase the efficiency of estimation and reduce the intercorrelations between estimates of the three independent parameters (Fisher, 1950; Kendall, 1951; Kendall & Stuart, 1973). The information function is used for this purpose.[1] The corrected approximations are the first stage ancillary estimators.

The corrections, $c(.)$, to the approximations for the three parameters are given by

$$(13)\quad c(\hat{\hat{a}}_i) = \left\{ I(\hat{\theta})^{-1} \left[\frac{\partial I(\hat{\theta})}{\partial \hat{\hat{a}}_i} - \frac{\partial^2 I(\hat{\theta})}{\partial \hat{\hat{a}}_i \partial \hat{\hat{c}}_i} \right] \right\} \hat{\theta} = \hat{\hat{b}}_i$$

---

[1] *Since* $-\frac{\partial^2 log\ \phi(\theta)}{\partial \theta^2}$ *or unity summarizes the information from an assumed prior of* $N(0,1)$, *the information function in this context is one plus the usual logistic information function as given by Birnbaum (1968).*

for item discriminatory power, by

$$(14)\quad c(\hat{\hat{b}}_i) = \left\{ I(\hat{\theta})^{-\frac{1}{2}} + I(\hat{\theta})^{-1} \left[\frac{\partial I(\hat{\theta})}{\partial \hat{\hat{a}}_i} + \frac{\partial I(\hat{\theta})}{\partial \hat{\hat{c}}_i} \right] \right\}_{\hat{\theta} = \hat{\hat{b}}_i}$$

for item difficulty, and by

$$(15)\quad c(\hat{\hat{c}}_i) = \left\{ I(\hat{\theta})^{-1} \left[\frac{\partial I(\hat{\theta})}{\partial \hat{\hat{a}}_i} - \frac{\partial^2 I(\hat{\theta})}{\partial \hat{\hat{a}}_i \partial \hat{\hat{b}}_i} \right] \right\}_{\hat{\theta} = \hat{\hat{b}}_i}$$

for the item coefficient of guessing. Equations (13) through (15) reduce to

$$(16)\quad c(\hat{a}_i) = \left[I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i} \right]^{-1} \left[\frac{1.445\hat{\hat{a}}_i}{(1 + \hat{\hat{c}}_i)^2}(3 - \hat{\hat{c}}_i^2) \right]$$

$$(17)\quad c(\hat{\hat{b}}_i) = \left[I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i} \right]^{-\frac{1}{2}} + \left[I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i} \right]^{-1} \left[\frac{1.445\hat{\hat{a}}_i}{(1 + \hat{\hat{c}}_i)^2}(1 - \hat{\hat{a}}_i - \hat{\hat{c}}_i^2) \right]$$

and

$$(18)\quad c(\hat{\hat{c}}_i) = \left[I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i} \right]^{-1} \left[\frac{1.445\hat{\hat{a}}_i}{(1 + \hat{\hat{c}}_i)^2}(2 - \hat{\hat{a}}_i) \right].$$

The ancillary estimators, $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, are obtained through

$$(19)\quad \hat{a}_i = \hat{\hat{a}}_i + c(\hat{\hat{a}}_i)$$

$$(20)\quad \hat{b}_i = \hat{\hat{b}}_i + c(\hat{\hat{b}}_i)$$

and

$$(21)\quad \hat{c}_i = \hat{\hat{c}}_i + c(\hat{\hat{c}}_i) \quad .$$

Notice that the corrections are negligible if the information function at $\hat{b}_i$, $I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i}$, is substantial. For example, as $I(\hat{\theta})_{\hat{\theta}=\hat{\hat{b}}_i}$ becomes infinite the outputs of equations (16), (17), and (18) approach zero. The implication of this feature is that there is no no need for corrections if the latent trait is manifest or, synonymously, if the ICC's have been fitted to θ, latent ability, as opposed to $\hat{\theta}$, manifest ability.

The procedure continues with certain refinements. Instead of using the standardized adjusted score as a substitute for latent ability, as indicated in equation (9), we may now use Bayesian modal estimates of ability (Samejima, 1969) in substitution. In general, these estimates of ability bear a relationship to latent ability that is more linear in form than that of standardized test scores. The estimates $\hat{P}_i'$ and $\hat{\delta}_i'_{\hat{\theta}}$ are computed and the Bayesian modal estimates are grouped into k intervals $[-\infty, -2.45], [-2.45, -2.40] \ldots [2.45, \infty]$ where k is equal to 100. The sample estimates, $\hat{P}_i'$ and $\hat{\delta}_i'_{\hat{\theta}}$, are then used as before with provisional $\hat{\hat{c}}_i$ to find the minimum of

$$(22)\quad \chi_i^2 = \sum_{j=1}^{k} \left[r_j - n_j P_i(\mu_j) \right]^2 \left[n_j P_i'(\mu_j) Q_i'(\mu_j) \right]^{-1}$$

where μ_j is the rescaled midpoint of interval j, for $j = 1, 2, \ldots k$. The midpoints are rescaled to be consistent with mean of zero and a standard deviation of one. The discussion attendant to finding the minimum with equation (9) is still appropriate here. We have merely replaced j with μ_j. When the minimum for equation (22) is found, the particular values of $\hat{\hat{a}}_i$, $\hat{\hat{b}}_i$, and $\hat{\hat{c}}_i$ providing that minimum are the final stage approximations. The corrections given in equations (16) through (18) are recalculated and applied as indicated in equations (19) through (21) to yield the final stage ancillary estimators.

This procedure has been employed at the U.S. Civil Service Commission to obtain estimates of the parameters for some 700 distinct items (Urry, 1974b). These items had been used in fifteen forms of a 60-item test of verbal ability. The samples used to parameterize the items in each of the fifteen forms ranged in size from 2,076 to 3,902 cases with an average of 3,072 cases per form. It will be recalled that the procedure involves the calculation of Bayesian modal estimates of ability for all cases. Thus the composite distribution of the estimates of ability can be examined with regard to normality as assumed by the procedure. A histogram is provided in Fig. 1 representing the distribution of ability estimates for the 46,079 cases involved in this large scale parameterization study. Visual inspection of this histogram indicates that the distribution of ability in this study was

reasonably well approximated by a normal distribution.

At the U.S. Civil Service Commission, we have been and currently are using the best 200 items from this parameterization study to tailor tests to live examinees. A report of the results of this research will be available early next year.

REFERENCES

Birnbaum, A. Part 5. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M., and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Fisher, R. A. *Contributions to mathematical statistics.* New York: John Wiley & Sons, 1950.

Kendall, M. G. *The advanced theory of statistics*, Vol. 2. (2nd ed.) London: Charles Griffin, 1951.

Kendall, M. G., and Stuart, A. *The advanced theory of statistics*, Vol 2. (3rd ed.) New York: Hafner, 1973.

Lord, F. M., and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 1969, No. 17.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, 34, 253 - 269. (a)

Urry, V. W. *Computer-assisted testing: calibration and evaluation of the verbal ability bank.* (TS-74-3) Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, December, 1974. (b)

THE URRY ITEM PARAMETER ESTIMATION TECHNIQUE: HOW EFFECTIVE?

Frank L. Schmidt
John F. Gugel[1]

Personnel Research and Development Center
U.S. Civil Service Commission

The purpose of this paper is to report the results of two simulation studies evaluating the effectiveness of the Urry item parameterization technique (Urry, 1975). The first study examines the performance of the technique under "standard" parameterization conditions; the second explores the limits of the technique's accuracy as number of subjects and number of items decrease.

Study I: Item parameterization under standard conditions.

METHOD

The items parameterized in this study are shown in Table 1. Three thousand three hundred simulated subjects were drawn randomly from $\theta = N(0,1)$ and, using the known item parameters, responses (0 or 1) of each subject to each item were generated using the process described in Appendix A. This process produced the basic data to which the parameterization procedure was applied. The procedure was applied to two samples -- the first 2,000 and 3,000 cases generated. Correlations were computed between estimated and known item parameters for all three item parameters in both samples and at various stages of parameterization. The root-mean-square error was also computed for each parameter using the following formula:

$$RMSE = \Sigma \{(P - \hat{P})^2 \, n^{-1}\}^{\frac{1}{2}} ,$$

where P = known values of a_i, b_i, or c_i, $\hat{P}$ = the estimated parameter values, and n = the number of items, here 100.

One hundred of the remaining 300 simulated examinees were used in each of three tailored testing conditions: the item parameters (1) were known, (2) had been estimated in the 2,000 case sample, and (3) had been estimated in the 3,000 case sample. The elegant Bayesian algorithm provided by Owen (1969) was employed in the tailored testing. This adaptive algorithm allows termination of tailored testing when any desired degree of accuracy has been attained. The measure of accuracy is the standard error of estimate of ability, which is routinely calculated after each item response is scored. Eight termination rules were used in this study, ranging from .5477 (corresponding to a reliability of .70) to .2366 (corresponding to a reliability of .95). Fidelity coefficients, $r_{\hat{\theta}\theta}$, were computed for combinations of termination criteria and conditions, as was the mean number of items required to reach the termination rule. These two indices were then compared across the three tailored testing conditions.

RESULTS

---

[1]*The computer processing was done through the University of Maryland Computer Science Center in conjunction with graduate work by the second author. Arrangements for computer time were made by faculty advisor Professor Charles E. Johnson of the Department of Measurement and Statistics, College of Education, University of Maryland.*

TABLE 1

True Parameters of the 100 Items Parameterized
(Study I)

Item	Parameters			Item	Parameters		
(i)	a_i	b_i	c_i	(i)	a_i	b_i	c_i
1	.80	-1.90	.03	51	1.60	.10	.18
2	.80	-1.70	.06	52	1.60	.30	.21
3	.80	-1.50	.09	53	1.60	.50	.24
4	.80	-1.30	.12	54	1.60	.70	.27
5	.80	-1.10	.15	55	1.60	.90	.03
6	.80	-.90	.18	56	1.60	1.10	.06
7	.80	-.70	.21	57	1.60	1.30	.09
8	.80	-.50	.24	58	1.60	1.50	.12
9	.80	-.30	.27	59	1.60	1.70	.15
10	.80	-.10	.03	60	1.60	1.90	.18
11	.80	.10	.06	61	2.00	-1.90	.21
12	.80	.30	.09	62	2.00	-1.70	.24
13	.80	.50	.12	63	2.00	-1.50	.27
14	.80	.70	.15	64	2.00	-1.30	.03
15	.80	.90	.18	65	2.00	-1.10	.06
16	.80	1.10	.21	66	2.00	-.90	.09
17	.80	1.30	.24	67	2.00	-.70	.12
18	.80	1.50	.27	68	2.00	-.50	.15
19	.80	1.70	.03	69	2.00	-.30	.18
20	.80	1.90	.06	70	2.00	-.10	.21
21	1.20	-1.90	.09	71	2.00	.10	.24
22	1.20	-1.70	.12	72	2.00	.30	.27
23	1.20	-1.50	.15	73	2.00	.50	.03
24	1.20	-1.30	.18	74	2.00	.70	.06
25	1.20	-1.10	.21	75	2.00	.90	.09
26	1.20	-.90	.24	76	2.00	1.10	.12
27	1.20	-.70	.27	77	2.00	1.30	.15
28	1.20	-.50	.03	78	2.00	1.50	.18
29	1.20	-.30	.06	79	2.00	1.70	.21
30	1.20	-.10	.09	80	2.00	1.90	.24
31	1.20	.10	.12	81	2.40	-1.90	.27
32	1.20	.30	.15	82	2.40	-1.70	.03
33	1.20	.50	.18	83	2.40	-1.50	.06
34	1.20	.70	.21	84	2.40	-1.30	.09
35	1.20	.90	.24	85	2.40	-1.10	.12
36	1.20	1.10	.27	86	2.40	-.90	.15
37	1.20	1.30	.28	87	2.40	-.70	.18
38	1.20	1.50	.06	88	2.40	-.50	.21
39	1.20	1.70	.09	89	2.40	-.30	.24
40	1.20	1.90	.12	90	2.40	-.10	.27
41	1.60	-1.90	.15	91	2.40	.10	.03
42	1.60	-1.70	.18	92	2.40	.30	.06
43	1.60	-1.50	.21	93	2.40	.50	.09
44	1.60	-1.30	.24	94	2.40	.70	.12
45	1.60	-1.10	.27	95	2.40	.90	.15
46	1.60	-.90	.03	96	2.40	1.10	.18
47	1.60	-.70	.06	97	2.40	1.30	.21
48	1.60	-.50	.09	98	2.40	1.50	.24
49	1.60	-.30	.12	99	2.40	1.70	.27
50	1.60	-.10	.15	100	2.40	1.90	.03

Correlations between estimated and known parameters are summarized in Table 2 for both sample sizes and by parameterization stages. Correlations tend to increase across parameterization stages. For example, the correlation, $r_{\hat{a}a}$ for the 3,000 case sample progressively increases stagewise from a low of .884 for approximations based on the regression of the binary items on the adjusted and standardized raw score distribution, to .895 for ancillary estimates obtained by correcting these approximations, to .914 for approximations based on the regression of the binary items on the distribution of Bayesian modal estimates of ability, to a high of .918 for ancillary estimates obtained by correcting the last approximations. Similarly, the $r_{\hat{b}b}$ and $r_{\hat{c}c}$ illustrate an increase by parameterization stages for the 3,000 case sample. For the 2,000 case parameterization sample, the same stagewise interpretation may be given to the $r_{\hat{a}a}$, $r_{\hat{b}b}$, and $r_{\hat{c}c}$. Increases in the correlations occur as the item analytic procedure continues to completion. On the basis of this analysis, 2,000 cases appear adequate for the parameterization procedure. Only negligible differences in correlations occur across sample sizes.

Root-mean-square errors are shown in Table 3 for both samples and by parameterization stages. Notice that these measures of error are reduced as the item analytic procedure continues. The only exception is in the case of b_i for final stage ancillary estimates in the 3,000 case sample. Again 2,000 cases appear adequate for the parameterization procedure on the basis of this analysis.

The results of the simulated tailored testing conducted to evaluate the efficiency of the parameter estimates are presented in Table 4. The eight termination rules are numbered in column 1. Column 2 shows the standard error of estimate for each rule, and column 3 shows the corresponding reliability coefficient, determined using the relation:

$$\rho_{\hat{\theta}\theta}^2 = 1 - \sigma_\varepsilon^2$$

The square root of the reliability coefficient is $\rho_{\hat{\theta}\theta}$, the fidelity coefficient, and is shown in column 4. The expected fidelity coefficients in column 4 can be compared to the obtained coefficients in columns 5, 7, and 9, which were gotten by actually correlating $\hat{\theta}$ with θ at each termination rule-condition combination.[2] At all termination rules, the performance of the parameter estimates based on the 2,000 and 3,000 case samples is virtually indistinguishable from that of the known or true parameters. The mean number of items required to reach the termination rule also differs only negligibly across conditions. (The number of items required is actually smaller in the case of the parameter estimates; we shall return to this phenomenon in Study II.)

The conclusion, then, is that, at least under the conditions examined in this study, the Urry item parameterization technique produces parameter estimates that:

1. are highly correlated with true parameter values,
2. show only small average squared deviations from true parameter values, and
3. perform virtually as well as the true parameter values when used in testing.

In addition, it appears that each of the steps in the parameterization procedure contributes to the accuracy of the final estimates and that 2,000 cases and 100

---

[2] *The reader may wonder whether the distinction between the fidelity coefficients in columns 4 and 5 (Table 4) is not academic. In fact it is not. The coefficients in column 4 are based solely on the pre-set termination rule and represent expected fidelity levels given a perfect item bank. The item bank examined here is not perfect, so even though the item parameters are known, (i.e., we have errorless parameter estimates), the coefficients in column 5 are somewhat smaller than those in column 4. A perfect bank is one with an infinite information function across the range of θ values from $+\infty$ to $-\infty$. In practice, information function values are not high beyond $\theta = \pm 2.5$, since it is difficult to locate, parameterize, and maintain sufficient items in this range. The result is a higher standard error of estimate for $\hat{\theta}$ in the extremes, and fidelity coefficients that are slightly less than maximal. Thus the process here is a weak form of the classical attenuation paradox.*

TABLE 2

Correlations Between Estimates and Known Parameters by Parameterization Samples and Stages (Study I)

Sample Size	Parameterization Stage	Correlation		
		$r_{\hat{a}a}$	$r_{\hat{b}b}$	$r_{\hat{c}c}$
2000	Corrected Raw Score:			
	Approximation	.876	.996	.651
	Ancillary Estimate	.873	.996	.668
	Bayesian Modal:			
	Approximation	.909	.996	.754
	Ancillary Estimate	.915	.996	.764
3000	Corrected Raw Score:			
	Approximation	.884	.996	.611
	Ancillary Estimate	.895	.996	.616
	Bayesian Modal:			
	Approximation	.914	.997	.752
	Ancillary Estimate	.918	.997	.760

TABLE 3

Deviation Measures for Estimates by Parameterization
Samples and Stages
(Study I)

Sample Size	Parameterization Stage	Deviation Measure $\left(\sum_{i=1}^{m} \{\hat{a}_i - a_i\}^2 m^{-1}\right)^{\frac{1}{2}}$	$\left(\sum_{i=1}^{m} \{\hat{b}_i - b_i\}^2 m^{-1}\right)^{\frac{1}{2}}$	$\left(\sum_{i=1}^{m} \{\hat{c}_i - c_i\}^2 m^{-1}\right)^{\frac{1}{2}}$
2000	Corrected Raw Score:			
	Approximation	.309	.181	.077
	Ancillary Estimate	.283	.120	.067
	Bayesian Modal:			
	Approximation	.269	.150	.061
	Ancillary Estimate	.242	.123	.056
3000	Corrected Raw Score:			
	Approximation	.308	.139	.081
	Ancillary Estimate	.253	.135	.073
	Bayesian Modal:			
	Approximation	.252	.109	.059
	Ancillary Estimate	.228	.148	.056

TABLE 4

Fidelity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for Tailored Testing to Various Termination Rules Where the Item Parameters Were Known or Estimated

#	Termination Rules $\sigma_{\hat{\theta}}$	$\rho^2_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	Parameters Known		Items Parameters Estimated in a Sample of: 2,000 Cases		3,000 Cases	
				$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	.5477	.70	.84	.84	2.7	.83	2.0	.84	2.3
2	.5000	.75	.87	.85	3.2	.86	2.7	.86	2.6
3	.4472	.80	.89	.89	3.9	.89	3.4	.88	3.2
4	.3873	.85	.92	.91	4.7	.90	4.0	.90	4.0
5	.3162	.90	.95	.94	6.6	.92	5.4	.93	5.6
6	.2828	.92	.96	.96	8.2	.94	6.7	.93	7.1
7	.2449	.94	.97	.96	10.8	.95	9.1	.94	9.6
8	.2236	.95	.97	.96	13.3	.95	11.1	.95	11.9

items are adequate to produce accurate parameter estimates.

Study II: Effectiveness with Reduced Numbers of Items and Cases

Study I employed large sample sizes (N = 2000 and 3000 cases) and a relatively large number of items (n = 100). Under these conditions, we have seen that the procedure produces very accurate parameter estimates. We will now examine the effects of reduced numbers of cases and items on error in the parameter estimates and on the accuracy of tailored testing using those estimates. It is known *a priori*, of course, that reduction in either the number of cases or the number of items will, other things constant, tend to increase estimation errors. But it cannot be known in advance how large or practically significant such increases will be. The present study, exploratory in nature, is addressed to these questions.

METHOD

Based on suggestions by Lord (1968, p. 1016) and the results of Study I, it was decided to allow the number of items to vary in the range from 50 to 100 and the number of cases to range from 500 to 2000. The initial 100-item bank, from which the smaller banks were later selected, was characterized by a_i values ranging uniformly from .80 to 2.20,[3] b_i values distributed uniformly from -1.9 to +1.9, and c_i values from .02 to .24, also uniform in distribution. These parameter values are not different from what one might reasonably expect to find empirically given prescreening of items (Urry 1974; Jensema, 1972). In the reduced item samples, the a_i values were chosen in equal steps from .80 to 2.20. For example, there were five levels of a_i for the 50-item test and ten for the 100-item test. Ten values of b_i in equal steps between -1.9 and 1.9, inclusive, were arranged within each level of a_i. (An exception was the 55-item test, which had eleven values of b_i in equal steps between -1.9 and 1.9, inclusive, within each of its a_i values.) Values of a_i on different tests were matched with the same b_i values. The c_i values ranged from .02 to .24 in equal steps, irrespective of a_i and b_i. Values of θ, representing simulated subjects, were again sampled randomly from $N(0,1)$. Then the simulation procedure described earlier (See Appendix A) was again used to generate a vector of responses (1 = correct; 0 = incorrect) for each θ for the item bank in question using the known item parameters. Parameter estimation was then carried out using this simulated data. The two indices used to evaluate the parameter estimates were the same as in Study I.

Tailored testing using the Owen algorithm was carried out on 100 simulated subjects using first the known item parameters and then item parameter estimates obtained on 1000 cases and 60 items. To increase the number of items used in tailored testing to a more realistic level, another identical set of 60 items was parameterized on a separate, independent group of 1000 simulated subjects, and these "items" were combined with the original 60 to produce a bank with 120 items. In the case of the known parameters, both 60-item sets were entered into the tailored testing bank. The known parameters in this bank were used to generate the response vectors of the 100 simulated subjects, and these vectors, in turn, were used in the tailored testing.

RESULTS

Results produced by the parameterization

---

[3] *The choice of a_i values between .80 and 2.20 was indicated by item calibration experiences with the USCSC Test 700. It was found that values in this range occur frequently enough to provide sufficient items for test construction.*

procedure for varying combinations of sample size and number of items are shown in Tables 5 and 6. In both tables, the raw score approximations prior to application of the ancillary correction procedure are shown, as well as the final estimates, i.e., the Bayesian modal estimates with ancillary corrections applied. "Lost items" are those for which the estimation procedure did not converge because of insufficient cases in the tails of distribution. (There were no such failures of convergence in Study I.)

Looking at the RMSE's for the final estimates in Table 5, it can be seen that, in general, decreasing both sample size and number of items results in increased RSME's. This effect appears to be more pronounced for a_i than for the other two parameters. Moving from 50 to 60 items (sample size constant) appears to produce marked reductions in error for a_i, but beyond this, improvements in accuracy with increases in number of items are smaller. The b_i and c_i were estimated rather accurately throughout the range of both independent variables, although variation in sample size and number of items did have the expected effect. The last column in Table 1 reveals a tendency for items to begin to fail to converge during parameter estimation when sample size is dropped as low as 500. Sample size appears more crucial in this respect than number of items. Correlations between final parameter estimates and actual parameters, shown in Table 6, also pattern themselves as expected, within the limits of sampling error.[4]

Results of simulated tailored testing using known parameters and parameters estimated on a sample of 1000 with 60 items are shown in Table 7. The organization of this table is the same as that of Table 4 of Study I. Fidelity coefficients in column 4 are theoretical values based solely on the termination rule chosen. Those in column 5 were obtained by correlating the $\hat{\theta}$ produced using the known item parameters with known θ. As expected, they are essentially identical to the predicted theoretical validities. Those in column 7 were obtained by correlating the $\hat{\theta}$ produced using the parameter estimates with the known θ. As expected, they are somewhat lower than those in columns 4 and 5, but it can be noted that, as the termination rule becomes more stringent, the discrepancy decreases. At the most stringent termination rule, the validity of the $\hat{\theta}$ derived using the parameter estimates is only .025 lower than that based on the known parameter. The reliabilities of the two $\hat{\theta}$'s at this termination rule are .95 and .90, respectively.

Why are the termination rules not fully attained when the parameter estimates are used? The tailoring algorithm capitalizes on an error in the parameter estimates. As a consequence, tailored testing using the estimated parameters terminates prior to actually reaching the pre-set termination rule. That is, because of capitalization on error in parameter estimates during the process of item selection, the reliability levels computed by the Owen algorithm at any stage during the tailored process are somewhat inflated. This leads to a too early termination of tailored testing, and, when the obtained $\hat{\theta}$ are correlated with θ, it becomes evident that the pre-set reliability level for termination has not been met. In the present example, an average of 14.57 items was administered when the known parameters were used but only 11.12 when the parameter estimates were used. This shrinkage problem can be overcome by setting the

---

[4]In examining these correlations, one should bear in mind that in the case of $\hat{a}_i$ and to a lesser extent $\hat{c}_i$, restriction in range is operating to lower the tabled values. The items parameterized contained no values of a_i lower than .80. This value of a_i corresponds to a biserial correlation of .62 between the item and latent ability. Past studies (Jensema, 1972; Urry, 1974b) have shown that only about one-third of the items in conventional tests have a_i values this large. No c_i greater than .24 were included; in practice c_i does exceed .24, although the range restriction here is probably less severe than in the case of a_i. The rather high a_i values among the items parameterized must also be considered in evaluating the tabled root mean squared errors for a_i. Errors in a_i are much larger for high a_i than low a_i, since when a_i is high, small errors in $\hat{\rho}_{I\theta}$ lead to large errors in $\hat{a}_i$. For example, if $\rho_{I\theta} = .90$, $a_i = 2.01$. If $\hat{\rho}_{I\theta} = .88$, $a_i = 1.85$, a difference of .16. But if $\rho_{I\theta} = .50$, $a_i = .58$. Then if $\hat{\rho}_{I\theta} = .48$, $\hat{a}_i = .55$ a difference of only .03. These same considerations apply, of course, to the results of Study I.

Table 5

Root Mean Square Errors of Corrected and Uncorrected Parameter Estimates
(Study II)

Items	Cases	Raw Score Approximations			Final Ancillary Estimates			Lost Items
		a_i	b_i	c_i	a_i	b_i	c_i	
50	2000	.283	.124	.086	.395	.137	.064	0
50	1000	.292	.193	.097	.326	.209	.078	1
50	500	.370	.164	.097	.472	.259	.077	0
55	2000	.385	.195	.091	.308	.150	.057	0
55	1000	.352	.194	.101	.315	.124	.071	0
55	500	.281	.185	.098	.403	.227	.086	4
60	2000	.321	.204	.091	.253	.140	.065	0
60	1000	.343	.231	.089	.322	.144	.062	0
60	500	.360	.194	.080	.342	.179	.068	0
70	2000	.272	.131	.095	.225	.166	.067	1
70	1000	.324	.189	.095	.273	.174	.074	1
70	500	.386	.197	.096	.351	.187	.083	4
80	2000	.266	.141	.092	.214	.150	.072	1
80	1000	.259	.178	.092	.261	.166	.073	1
80	500	.319	.224	.091	.311	.229	.079	6
90	2000	.297	.180	.094	.244	.149	.069	0
90	1000	.341	.171	.089	.304	.140	.072	0
90	500	.316	.184	.094	.283	.144	.086	2
100	2000	.290	.138	.085	.223	.131	.056	0
100	1000	.286	.137	.088	.240	.162	.062	0
100	500	.354	.189	.100	.276	.161	.083	5

$$RMSE = \sqrt{\frac{\Sigma(p_i - \hat{p}_i)^2}{n}}$$

Where p = parameters
n = number of items

Table 6

Correlation Between Parameter Estimates and Known Parameters
(Study II)

Items	Cases	Raw Score Approximations $r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$	Final Ancillary Estimates $r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$
50	2000	.846	.999	.599	.849	.997	.636
50	1000	.888	.992	.429	.908	.990	.492
50	500	.745	.993	.428	.780	.989	.454
55	2000	.731	.995	.488	.891	.995	.646
55	1000	.758	.995	.428	.870	.995	.546
55	500	.850	.992	.387	.824	.990	.376
60	2000	.828	.996	.491	.899	.997	.630
60	1000	.771	.994	.546	.842	.995	.588
60	500	.768	.994	.626	.801	.995	.668
70	2000	.834	.997	.471	.922	.997	.632
70	1000	.813	.996	.468	.828	.996	.521
70	500	.715	.993	.464	.813	.995	.449
80	2000	.873	.996	.535	.914	.997	.574
80	1000	.850	.994	.465	.879	.993	.550
80	500	.839	.991	.410	.823	.989	.502
90	2000	.861	.996	.483	.871	.996	.568
90	1000	.757	.995	.518	.847	.995	.547
90	500	.804	.995	.447	.874	.993	.418
100	2000	.837	.997	.539	.877	.998	.690
100	1000	.843	.996	.470	.863	.996	.627
100	500	.741	.993	.344	.824	.994	.420

Table 7

Fidelity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for Tailored Testing to Various Termination Rules Where the Item Parameters Were Known or Estimated
(Study II)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Termination Rules			Parameters Known		Parameters Estimated	
#	σ_ε	$\rho_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
1	.5477	.70	.84	.864	2.43	.792	2.26
2	.5000	.75	.87	.904	3.31	.821	2.89
3	.4472	.80	.89	.932	4.00	.821	2.89
4	.3873	.85	.92	.935	4.91	.864	3.70
5	.3162	.90	.95	.955	7.03	.895	5.30
6	.2828	.92	.96	.962	8.77	.921	6.57
7	.2449	.94	.97	.969	11.77	.942	8.91
8	.2236	.95	.97	.975	14.51	.952	11.12

reliability termination rule higher than that actually required. In our present example, the termination rule should be set at .95 in order to obtain $\hat{\theta}$ of reliability .90. However, parameterizing on larger sample sizes (both number of items and numbers of cases) and thus obtaining more accurate initial estimates is preferable to adjusting termination rules to allow for shrinkage. As we saw in Study I, a parameterization sample of 2000 cases and 100 items is sufficient to virtually eliminate the shrinkage problem. The results of Study II indicate that a parameterization sample of 1000 cases and 60 items is not sufficient in this respect.

DISCUSSION

The results obtained in these two studies indicate that the Urry item parameterization procedure works quite well under what might be referred to as "standard" conditions. Given at least 2000 cases and 100 items of good but not unrealistically high quality, the procedure produces estimates that correlate highly with true parameters, show low RMSE's, and perform about as well when used in tailored testing as the true parameter values. As number of cases and/or number of items in the parameterization sample decreases, correlation of estimates with true parameters decrease and RMSE's increase. When these parameter estimates are used in tailored testing, capitalization on error by the Owen algorithm[5] during item selection results in termination of tailored testing prior to attaining the pre-set criterion of score reliability. This condition is quite noticeable at most termination rules when 60 items and 1000 cases are used in the parameterization sample. Whenever parameter estimates fully capable of meeting termination rules are necessary or desired, we recommend that the parameterization procedure be used with sample sizes in the neighborhood of 2000 (or greater) and not less than 100 items. On the other hand, when one is willing and able to compensate for shrinkage by increasing the stringency of the termination rule, parameter estimates based on 1000 cases and 60 items will probably be adequate.

[5] *Any other adaptive algorithm would show the same effect.*

REFERENCES

Jensema, C.J. *An application of latent trait mental test theory to the Washington pre-college testing battery.* Unpublished doctoral dissertation. University of Washington, 1972.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement,* 1968, 28, 989-1020.

Urry, V. W. *Computer-assisted testing: the calibration and evaluation of the verbal ability bank.* (TS-74-3) Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Paper presented at the Symposium, *Computers and testing: steps toward the inevitable conquest.* APA Convention, Chicago, Illinois, August 31, 1975.

APPENDIX A

Model Sampling and Simulation Techniques

Let us assume that latent ability is normally distributed. As a consequence, the jth subject may be simulated by randomly selecting a value of Θ_j from the distribution N(0,1). For the ith item having known parameters, the jth simulated subject's binarily scored response, i.e., 0=wrong, 1=right, is obtained through

$$P_i^{(\Theta_j)} = c_i + (1 - c_i) \; P_i(\Theta_j) \; , \tag{1}$$

where $P_i^{(\Theta_j)}$ = the proportion answering the item correctly. $P_i^{(\Theta_j)}$, the proportion actually knowing the correct answer is:

$$P_i(\Theta_j) = \frac{1}{+ \exp\{-D a_i (\Theta_j - b_i)\}} \tag{2}$$

Since

$$Q_i^{(\Theta_j)} = 1 - P_i^{(\Theta_j)}, \tag{3}$$

the exhaustive and mutually exclusive events 0 (wrong) and 1 (right) may be mapped onto the unit interval. A random number designated r_u is drawn from the distribution U(0,1). The following assignments are then appropriate:

$$u_{ij} = 0 \text{ (wrong), given } r_u > P_i^{(\Theta_j)} \tag{4}$$

$$u_{ij} = 1 \text{ (right), given } r_u < P_i^{(\Theta_j)}. \tag{5}$$

The jth subject's vector

$$v_j = (u_{1j}, u_{2j}, \ldots u_{mj}) \tag{6}$$

is generated for constant Θ_j, where m is the total number of items of known parameters. The subscript i must denote constant item parameters, when we model sample the v_j for $j = 1, 2, \ldots N$. Notice that the independent and random draws from $u(0,1)$ which total mN are in strict accord with the assumption of the local independence required by the model and that the generated data conform strictly to the defined model.

DISCUSSION

Frederic M. Lord
Educational Testing Service

It is a really impressive fact -- I am very much impressed -- that the United States Civil Service Commission is successfully trying out tailored testing on an experimental basis. They feel that the economic costs of such testing can be carried -- or at least that this will be true by the time they are ready for full-scale operational testing. I do not see any other serious obstacle to their success. The theoretical and empirical results presented here today convince me of this.

Aside from hardware and computer software, which are certainly important, what are the other essentials for a successful tailored testing operation?

1. An accurate method for estimating item parameters.
2. A rule for selecting the items to be administered, so as to measure each individual as accurately as possible.
3. A method of scoring, so that examinees who answer different items can nevertheless be put on the same score scale.
4. A method for evaluating the accuracy of measurement actually achieved, for an individual, and over groups of individuals.

1. Estimating Item Parameters

In tailored testing, it is necessary to predict an examinee's probability of success on items that are different from those he has already attempted. This is necessary in order to select the items to be administered. We must also be able to describe the ability of each examinee on a scale that is independent of the set of items administered.

Neither of these things can be done by classical test theory. To do these things (and more) it is necessary to use item characteristic curve theory. At least when there is guessing, item characteristic curve parameters cannot be simply and directly estimated from the usual statistics available: that is, proportion of correct answers and item-test correlations, or even item intercorrelations.

If the ability of each examinee were known, it would be a simple matter for a computer to estimate item characteristic curve parameters. If the item parameters were known, it would be a simple matter to estimate the ability of each examinee. To get off the ground, we must estimate both item and people parameters simultaneously. This suggests, obviously, an iterative procedure in which we treat crude estimates of one set of parameters (the person parameters, for example) as known in order to obtain tentative estimates of the other set of parameters (the item parameters). Then turn this process around and treat the second set of parameters as known, so as to reestimate the first set of parameters. These two steps can then be iterated.

Urry's method described today represents two iterations (or four steps) of this description. I imagine Urry would carry out additional steps if there were reason to. The empirical results reported by Schmidt and Gugel may indicate that additional iterations are not necessary.

The maximum likelihood estimation procedure that I use for research purposes typically carries out many more than two iterations. This is perhaps part of the reason why it seems more expensive than Urry's method. In 15 recent complete runs of the maximum likelihood method, the number of iterations carried out ranged from 9 to 43. Typically, the maximum likelihood program continues until the changes in the item parameters are in the fourth decimal place. Such accuracy is really unnecessary for most

practical purposes. I do it this way so that when I get some peculiar result, I can be sure it is not due to an inaccuracy in the estimation procedure.

2. Item Selection

To determine which item to administer next, Owen's method, as explained by Jensema, computes for each item the expected standard error for the ability estimate that would result if that item were administered next. The item giving the smallest standard error is the one selected for administration. Whenever test items differ in discriminating power, as they always do in practice, an item-selection method that takes discriminating power into account will surely be better than a method (such as the so-called branching method) that selects items solely on the basis of item difficulty. Much of the advantage of tailored testing demonstrated by the speakers today comes from using the most discriminating items rather than from matching the difficulty of the item to the ability of the examinee.

An alternative method is to select an item for administration on the basis of the item information function. The selected item would be the item that provided the most information at the examinee's estimated ability level. The information function takes into account both the standard error of the estimate and its bias, if any. Maximizing information is equivalent to minimizing the standard error of the estimated ability provided ability is estimated without bias.

3. Scoring

Since each examinee takes a different set of items, it seems inevitable that the so-called ability scale of item characteristic curve theory must be used to provide individuals with comparable scores. The examinee's score will be our best estimate of his ability. This estimate is supposed to be a consistent estimate (roughly speaking, an asymptotically unbiased estimate) regardless of the set of items administered. ('Asymptotically unbiased' here means that the estimate will be unbiased if the test is made long enough. This is not much to claim for an estimate, but we would be quite unhappy if our estimates did not have this property.

It is worthwhile to talk about two main methods of estimating ability:

1. The use of maximum likelihood estimates.
2. Owen's use of the Bayesian posterior mean, described by Jensema. Owen assumes a normal distribution of ability.

Any discussion of Bayesian versus non-Bayesian statistics is a hazardous enterprise for a discussant. In spite of this, I feel fairly comfortable with the task, perhaps because I feel no strong commitment to either approach. I feel that ultimately an empirical Bayes approach, different from either of these discussed here, will be preferred.

It has been asserted that mental testing theory and practice provides an excellent field for the application of Bayesian statistics. It is very reasonable to assume that the examinees tested are actually -- or can be treated as if they were -- a random sample from some population. Each individual has an unknown ability parameter. It is highly reasonable and proper to conceptualize any statistical inference problem in this context as depending on a prior distribution of ability.

The catch is that we must use the right prior distribution, not a wrong one. We often have quite large samples -- often large enough to prove that the prior distribution we would like to use is wrong, and cannot fit the data.

I start from a position not often stated explicitly -- that the usual Bayesian approach will yield results superior to any classical approach whenever the Bayesian prior is correct or appropriate; the usual Bayesian approach will probably yield poorer results whenever the chosen prior is grossly incorrect or inappropriate. All this seems rather obvious.

Urry can feel very happy with the distribution of ability estimates shown in his Figure 1. His result justifies the normality assumption in the estimation methods he uses.

Since I have developed and habitually

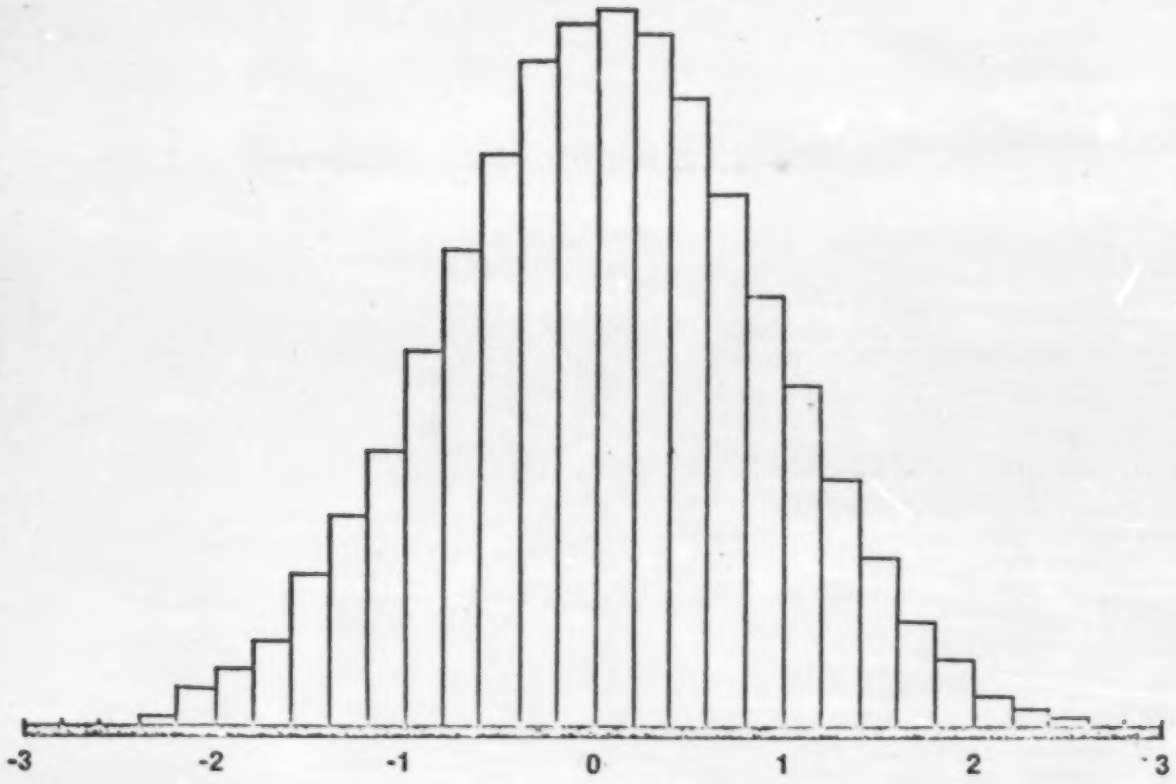


Figure 1. Composite histogram of Bayesian modal estimates of ability ($\hat{\theta}$) for 46,079 cases from fifteen parameterization samples.

use maximum likelihood estimates rather than Bayesian methods, it has to be my place to somehow try to fault Owen's Bayesian approach based on a normal prior. Yet, what can I say in the face of Figure 1, which (I have checked) is beautifully fitted by a normal distribution?

I wish my ability estimates, obtained by maximum likelihood, formed a normal distribution, like Urry's Figure 1. The histogram in my handout shows the distribution of 1,807 ability estimates for a representative group of examinees who answered the last item in the 60-item Mathematics section of the College Entrance Examination Board's *Scholastic Aptitude Test*. The smooth curve in the Figure is an estimate of the same distribution, obtained by a very different model that does not use the item responses but only the frequency distribution of number-right scores. The data and the models used are more fully described in my article in the June 1974 *Psychometrika*, from which this Figure was taken. The agreement shown between the distribution of estimated ability and the estimated distribution of ability is very good, giving me confidence in the effectiveness of both the models used.

I have looked at six different distributions of estimated ability obtained from ETS data. All are skewed to the left, but most of them only slightly so. If I had assumed a normal distribution, I imagine my distributions would be even more nearly normal than they are.

Perhaps the distribution in the handout is skewed because only those people who answered the last item in the test are included. But we cannot dismiss it on these grounds. If it is skewed for this reason, then we should expect other selected groups to have skewed distributions also. Surely, the highly selected group of college or graduate-school applicants should not be expected to have a symmetric distribution of ability.

I should call attention in passing to the extreme left of the histogram where a substantial number of cases appear with estimated abilities below -3. On every Scholastic Aptitude Test, there are always a number of people who are unlucky in their guesses and whose performance on the test is below the chance level -- that is, below the score level they would be expected to achieve by sheer random guessing. Such examinees show up with ability estimates of minus infinity when maximum likelihood is used.

It is a practical advantage of the maximum likelihood method that, for purposes of estimating item parameters, we can work with groups of examinees intentionally selected to have a roughly rectangular distribution of ability. Doing this greatly reduces the sampling error of many of the item-parameter estimates. This may be seen from the fact that the item characteristic curve is the regression of item score on ability. Regression curves are best estimated when there are many observations at the extremes of the data.

Let me now raise what may be an important theoretical and practical question. Suppose our item characteristic function $P_i(\theta)$ is logistic. In the usual notation,

$$(1)\quad P_i(\theta) = c_i + (1 - c_i)[1 + \exp\{-1.7a_i(\theta - b_i)\}]^{-1}.$$

Let us transform to a new ability scale, denoted by θ^*. Suppose

$$(2)\quad \theta^* \equiv \theta^*(\theta) \equiv Ke^{k\theta},\ b_i^* \equiv Ke^{kb_i},$$

$$a_i^* \equiv 1.7a_i/k,$$

where K and k are any positive constants. Now,

$$(3)\quad P_i(\theta) = c_i+(1-c_i)[1+e^{-1.7a_i\theta}e^{1.7a_ib_i}]^{-1}$$

$$= c_i+(1-c_i)[1+(b_i^*/\theta^*)^{a_i^*}]^{-1}$$

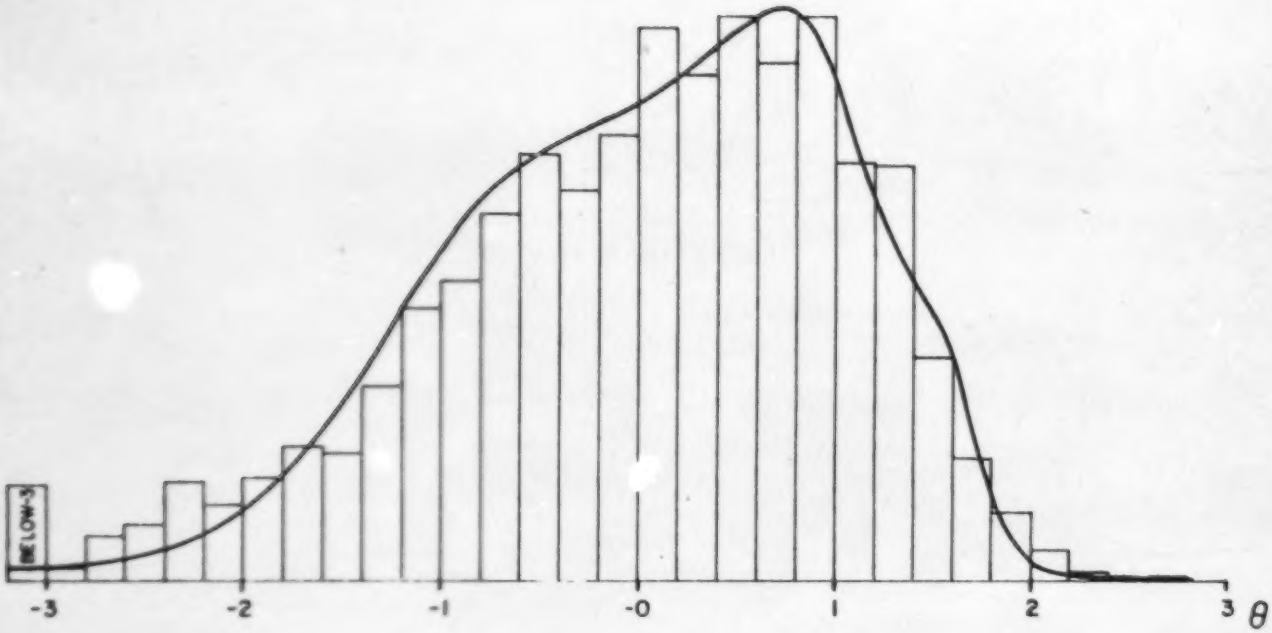
With this formulation,

$$(4)\quad Q_i \equiv 1 - P_i = (1 - c_i)\frac{(b_i^*/\theta^*)^{a_i^*}}{1 + (b_i^*/\theta^*)^{a_i^*}}$$

and

$$(5)\quad \frac{P_i - c_i}{Q_i} = \left(\frac{\theta^*}{b_i^*}\right)^{a_i^*}$$

This last equation relates the probability of success on an item to the simple ratio of examinee ability to item difficulty, an



Distribution of Estimated θ (histogram) and Estimated Distribution of θ (curve). Reproduced from Lord (1974).

intuitively helpful way of viewing the situation. The more ability you have the greater the probability of a correct response; the more difficult the item, the less the probability of a correct response.

I would argue that (3) may be a simpler and more natural form in which to express the model than the logistic form (1). I make this as a serious suggestion.

Perhaps it is θ^* that we should expect to have a prior distribution that is normal. If θ^* is normally distributed, then θ is not, and vice versa. The maximum likelihood estimate of a transformed θ is simply the transformation of the maximum likelihood estimate of θ:

$$\hat{\theta}^* = \theta^*(\hat{\theta}).$$

Thus a transformation of the ability scale makes no essential difference when maximum likelihood estimation is used.

Suppose next we assume, as Owen does, that ability is normally distributed with zero mean and unit variance. Let me raise one question which is not so serious as it may first appear. A person whose true ability is at +3 has no chance of being measured correctly under our assumption, since such an extreme person would occur only once out of 370 people, under our assumed prior. If the true distribution of ability is skewed, or high-tailed, there might be many people at +3 in our sample, yet none would be measured correctly under our incorrect prior.

This conclusion is reinforced by looking at Urry's Figure 1. Only about 330 people out of 46 thousand turn out to have estimated ability above +2, when there should be more than 1,000 according to the normal prior assumption. Only about 250 people turn out to have estimated ability below -2, when there should be more than 1,000 people. The reason, is that the normal prior makes extremes of ability hard to believe. Thus, the assumption of a normal prior guarantees that our estimates will be regressed towards the mean. The variance of the estimates of ability will be less than the variance of the true ability value. This is borne out by Urry's Figure 1. The standard deviation of the estimates of ability in Figure 1 is about .83, not 1.000.

The mean of the ability estimates in Figure 1 is about .05. Is there any importance in the fact that this is significantly nonzero?

The questions I have just raised may be of some importance if the estimated ability values are to be used for estimating item parameters, or for other exacting purposes. My objections are of no importance if the ability estimates are used simply to try to rank civil service applicants in order of merit. The rank order is probably not appreciably affected either by regression towards the mean, or by displacement of the mean. Thus as I see it, the CSC can confidently use Baysian estimates in real-life operational testing.